United States Environmental Protection Agency
Office of Water (4606)
www.epa.gov/safewater
EPA 816-R-00-020
October 2000

Data Reliability Analysis of the EPA Safe Drinking Water Information System / Federal Version

(SDWIS/FED)

Acknowledgements

The following people contributed to this project and the preparation of this report:

Jan Auerbach	Project lead
Lee Kyle	Data quality quantification, principal author of this report
Fran Haertel	Data quality characterization, state-specific data quality reports

Members of the Data Reliability Stakeholders Workgroup

EPA Headquarters:

Jan Auerbach, Chair; Chief, Information Management Branch (IMB)
Fran Haertel, Environmental Protection Specialist, IMB
Lee Kyle, Statistician, IMB
Ken Harmon, Office of Enforcement and Compliance Assistance (OECA)

EPA Regions:

1: Chris Ryan, Region I
2: Mark Rasso, Region II SDWIS/FED coordinator
5: Tom Poleck, Region V SDWIS/FED coordinator
6: Andy Waite, Region VI SDWIS/STATE coordinator
8: Aundrey Wilkins; Jack Rychecky, Region VIII Branch Chief

States:

Florida: Kenna Study, Drinking Water Manager
Iowa: Dennis Alt, Drinking Water Manager
Utah: Kevin Brown, Drinking Water Administrator
Washington: Peggy Johnson, Drinking Water Manager
Asssociation of State Drinking Water Administrators (ASDWA): Vanessa Leiby, Executive Director; Bob Blanco

Industry:

American Metropolitan Water Administrators: David Denig-Chakroff
American Water Works Association (AWWA): Teryl Pajor, Dan Schechter
National Association of Water Companies: John Hroncich

Other:

State Lab: Steve Jennis
Natural Resources Defense Council (NRDC): Eric Olson, Adriana Quintaro

EXECUTIVE SUMMARY

In 1998 EPA launched a major effort to assess the quality of its drinking water data—data used to assess compliance with the Safe Drinking Water Act, and found that it needs to be improved. This report is the culmination of that effort.

This report provides specific findings and estimates of the quality of the data that are in, and should be in, the EPA Safe Drinking Water Information System (SDWIS/FED). These results in no way should be interpreted as a reflection of drinking water quality, which overall remains high.

SDWIS/FED is EPA's drinking water database. It contains drinking water data for approximately 170,000 public water systems serving over 250 million people. Each water system has inventory data which describe the water system, data on any violations they have incurred, and resulting enforcement actions taken by states and/or EPA to ensure drinking water protection.

EPA found that the data quality for a selected subset of the required inventory data elements is high, that the data quality of violations data is low, and that enforcement actions data are of moderate data quality. SDWIS/FED data quality findings were found to be similar across water system types and size categories.

The violations listed in SDWIS/FED are accurate, but they are incomplete. A number of states have never reported certain types of violations. While industry found a few cases of over-reporting in the past (which have been corrected), EPA found very little over-reporting of violations in its analyses.

EPA and states have taken or scheduled a number of corrective actions, which are described in the Management Summary. These actions include (but are not limited to) more and improved training on rule implementation and data entry, additional and revised data audits, and improved data error interpretation. These corrective actions should improve the quality of the data reported nationally, as well as improve the public's understanding of the overall high quality of drinking water supplied by most systems in the United States.

PART I: MANAGEMENT SUMMARY

1 Introduction

In 1998, EPA launched a major effort to assess the quality of its drinking water data—data used to assess compliance with the Safe Drinking Water Act, and found that it needs to be improved. This report provides estimates of the quality of the data that are in, and should be in, the EPA Safe Drinking Water Information System (SDWIS/FED). These results in no way should be interpreted as a reflection of drinking water quality, which overall remains high.

SDWIS/FED is EPA's drinking water database. It contains drinking water data for approximately 170,000 public water systems serving over 250 million people. Each water system has inventory data which describe the water system, data on any violations they have incurred, and resulting enforcement actions taken by states and/or EPA.

EPA found that the data quality for a subset of 8 required inventory data elements is high, that the data quality of violations data is low (principally because they are incomplete), and that enforcement actions data are of moderate data quality.

1.1 Background

In the summer of 1998, some drinking water utility trade associations advised their members to check the EPA "Envirofacts" website, which contains violations and enforcement actions information on individual water systems. This was to prepare them for possible inquiries from their customers. Some larger utilities found gross errors in the reporting of violations against their water systems, specifically in cases of "over-reporting"— violations that never occurred which were listed in SDWIS/FED. Several of the utilities met with the incoming Assistant Administrator for the Office of Water, J. Charles Fox, to voice their concerns over the poor quality of the data that were available to the public.

Also that summer, EPA was preparing the first Annual Compliance Report (ACR), as required by the 1996 Amendments to the Safe Drinking Water Act. The Amendments required states to prepare state reports, and for EPA to compile them in a national report. When EPA compared the data in the state reports to the data these states submitted to SDWIS/FED, it found more than a 30% overall difference in data which should have been virtually identical. These two concerns led the Assistant Administrator to issue a letter to the states on September 3, 1998, calling for a major initiative to quantify and characterize the quality of the data in SDWIS/FED.

EPA began this initiative by holding three national public meetings on SDWIS/FED data quality in November and December 1998. EPA then formed a data reliability stakeholders workgroup comprised of people from EPA headquarters and regional offices, state drinking water programs, water utilities, industry associations, laboratories, and an environmental non-profit organization. The workgroup considered the comments from the public meetings and helped EPA develop a Data Quality Action Plan.

The Data Quality Action Plan, dated December 31, 1998 consisted of 4 major components:

1. Establish a SDWIS/FED data quality goal:

   "SDWIS/FED will contain 100% complete, accurate, timely, and consistent data which portray the data submitted by public water systems and primacy agencies, consistent with the Safe Drinking Water Act (SDWA) requirements. This goal will be advanced through interim milestones, which can be set once the current level of SDWIS/FED data quality is determined."
2. Improve the way SDWIS/FED data are presented in the EPA Envirofacts website.

   Several water utilities and other stakeholders raised concerns about what water system compliance information was available on the EPA Envirofacts website, and how it was displayed.
3. Take interim actions to improve SDWIS/FED data quality (status of these actions taken are discussed in Section 3.1).
4. Quantify and qualify the quality of SDWIS/FED data.

EPA used several analyses to quantify and characterize data quality. The overall SDWIS/FED data quality estimates for inventory, violations, and enforcement actions data are based primarily on the findings from the data verifications analysis, with input from an analysis comparing Annual Compliance Report (ACR) data to data in SDWIS/FED. The data verifications analysis included 29 data verification audits conducted in 27 states between 1996 and 1998. A total of 1,857 systems were audited (see Section 4.4 for details).

Initial SDWIS/FED data quality estimates were shared with states and EPA regions in summer 1999. Any errors found were checked and corrected. Some of the states concerns are discussed in Section 4.4.1.3 of this report. Since then, the other analyses have been completed, including the industry surveys. The overall findings are presented in this report.

Analyses used to quantify and characterize SDWIS/FED data quality

- Data verifications— reviews of data in state files that provided numerical estimates of overall SDWIS/FED data quality
- Industry surveys—water systems reviews of SDWIS/FED data that provided numerical estimates of the accuracy of data in SDWIS/FED
- Frozen database comparison—to develop numerical estimates of the timeliness in which violations are reported
- Comparison of SDWIS/FED data to Envirofacts data—to check for data transmission errors from one data set to the other
- Comparison of states' reporting of 1997 Annual Compliance Report (ACR) data to SDWIS/FED data—provided ratios of under-/over-reporting, reasons for discrepancies
- Errors analysis—to evaluate errors incurred in transferring data to SDWIS/FED

The outcome of the SDWIS/FED data quality analysis is to provide a benchmark of SDWIS/FED data quality, and a better understanding of where greater attention needs to be focused to improve it. The quantitative portion of this analysis also provides a water system perspective (% water systems having violations or enforcement actions) when available to provide additional perspective. The qualitative portion ties together numerical and non-numerical information from a number of analyses in an attempt to further characterize where the problems are occurring, and why.

1.2 Perspective/context

These results in no way should be interpreted as a reflection of drinking water quality, which overall remains high. Nor does it question the accuracy of the data submitted by laboratories or water systems to states (inaccurate lab results, fraud, data falsification, etc.). The thousands of compliance decisions that are made correctly by state drinking water programs are not enumerated. Only the violations and enforcement actions appear, because SDWIS/FED is an exceptions database (in other words, states do not provide sample data on regulated contaminants; they only report to SDWIS/FED when an “exception,” such as a violation or enforcement action, has occurred). As will be shown, only a small percentage of systems have any health-based violations. Many states have taken corrective steps to improve their SDWIS/FED data quality since these data were gathered.

2 Summary of findings

Summary estimates of SDWIS/FED data quality are presented in Text Box 1. Detailed estimates are contained in Part II of this report.

Inventory data

- The overall quality of 8 core SDWIS/FED inventory data parameters is high. That is, only 4% of the inventory parameters checked had any discrepancies (discrepancies are differences in data, missing data, or errors). The two parameters that change most frequently—population served and number of service connections—had the highest discrepancy rates. SDWIS/FED data quality estimates are very similar across water system types. These results are corroborated by the industry surveys.

8 inventory parameters checked:

System status (active or inactive)
Water system type
Primary source of water
Population served
# service connections
Address
Name
Water system ID

Violations data

- The overall quality of SDWIS/FED violations data is moderately high (estimated at 68%) for the Total Coliform Rule standard (an acute health-effects measure). However, it is very low for other health-based standards including Chemicals, Radionuclides, Surface Water Treatment Rule, and for monitoring/reporting requirements.
- Most of the discrepancies are because of unrecorded and unreported violations. This accounts for 56% of all MCL discrepancies, 83% of SWTR TT discrepancies, and 94% of all monitoring/reporting discrepancies. Data flow discrepancies (data in state databases but not SDWIS/FED) account for the remainder.

- The data that are reported in SDWIS/FED are highly accurate overall, in part because edit checks reject data which are transferred incorrectly.
- Data quality estimates are similar across water system types; this is corroborated by the industry surveys.
- Very little indication of over-reporting of violations was found (less than 0.7% of violation discrepancies).
- A number of states have never reported certain types of violations.
- Many states are not meeting the 90-day deadline for reporting violations. Only 68% of violations were reported on time.
- Violations reported using the Traditional method (selected data replacement or correction) appear to be more timely than those reported using the Total Replace method (replacing the entire data set each time changes are made).

Enforcement actions data

- SDWIS/FED enforcement actions data were found to be 87% complete and 83% accurate. Results were similar across water system types.

Other findings

- No discrepancies were found between data in SDWIS/FED and Envirofacts.
- “Data entry problems” was the most frequently cited reason for discrepancies between ACR data reported by states and SDWIS/FED data. “Resource limitations” was the next most common reason for discrepancies.
- Using the Traditional data entry method, 20% of inventory data and 32% of violations and enforcement actions data are being rejected. It was not possible to perform a similar calculation for the Total Replace method.
- Only 25% of all states were successful in resubmitting data in their first attempt. Of those not successful on the first attempt, 82% of the error types were data entry errors. Seven percent or less represent SDWIS/FED software limitations and problems.
- Characteristics of state programs that result in high quality SDWIS/FED data include routine, meaningful communication at all levels; Annual PWS notification of monitoring schedules; and automated monitoring compliance determination.

Text Box 1

SDWIS/FED Data Quality Summary Statistics

Inventory data

The SDWIS/FED data quality of 8 inventory parameters checked is estimated to be 96%.

Number of data points	16,006
Discrepancies:	
Number	646
Percent	4.0%
SDWIS/FED data quality	**96%**

~ 2,000 systems reviewed times 8 parameters checked

The # of instances where the DV audit team concluded that the parameter in SDWIS/FED was incorrect.

SDWIS/FED data quality = % of data without discrepancies (errors)

SDWIS/FED inventory data quality by parameter: System status (active or inactive)—97%, water system type—97%, primary source of water—98%, population served—91%, # service connections—92%, address—95%, name—98%, water system ID—100%.

Violations data

The SDWIS/FED data quality of violations data ranges from 7% for Surface Water Treatment Rule Treatment Technique (SWTR TT) violations to 68% for Total Coliform Rule Maximum Contaminant Level (TCR MCL) violations. Violations data listed in SDWIS/FED are accurate, but not incomplete. In addition, 68% of violations are reported on time.

	TCR MCL	Total Other MCL	SWTR TT	Total M/R
% systems w/ Violations	6.1%	< 4.3%	9.6%	< 78%
Number of Violations	162	59	94	5,091
Discrepancies				
Number	52	50	87	4,613
Percent	32%	85%	93%	91%
% Completeness	68%	19%	11%	10%
% Accuracy	99%	79%	67%	95%
SDWIS/FED data quality	**68%**	**15%**	**7%**	**9%**

The estimates on this line are not part of the data quality calculations, but lend perspective: 78% of the 1,857 systems reviewed had at least 1 violation of any type during the 1-3 year period of review for contaminants and rules

# violations that the DVs determined should have been reported to SDWIS/FED, whether or not they were

# errors cited in the DVs: 93% were for violations not designated by states as violations; remaining 7% occurred between state databases and SDWIS/FED

Completeness: % violations that should be in SDWIS/FED that made it in

Accuracy: % violations in SDWIS/FED that are correct

SDWIS/FED data quality = % of data without discrepancies (errors)

Enforcement actions data

The SDWIS/FED data quality of formal enforcement actions is estimated to be 72%. All formal enforcement actions, which are issued by the state and/or EPA in response to violations, are required to be reported to SDWIS/FED.

% systems with Enforcement Actions	20%
Number of Enforcement Actions	1,032
Discrepancies	
Number	287
Percent	28%
% Completeness	87%
% Accuracy	83%
SDWIS/FED data quality	**72%**

# errors cited in the DVs. The DV audits only measure the difference between state databases and SDWIS/FED. Auditors did not assume there should be an enforcement action unless the state actually took one

Completeness: % enforcement actions in state files that made it into SDWIS/FED

Accuracy: % enforcement actions in SDWIS/FED that are correct

SDWIS/FED data quality = % of data without discrepancies (errors)

3 Corrective actions

Text Box 2 defines the 4 elements of SDWIS/FED data quality and correlates their improvement to the corrective actions discussed below.

3.1 Early actions taken

Before waiting for the results of the analyses which would quantify and qualify the quality of the data in SDWIS/FED, the data reliability stakeholders workgroup recommended in December 1998, and EPA subsequently completed, several actions to improve SDWIS/FED data quality in the interim.

EPA HQ:

- Improved the way SDWIS/FED data are presented in the EPA Envirofacts website.

  In response to concerns about the quality of older SDWIS/FED data, now only violations and enforcement actions incurred since 1993 will be displayed in Envirofacts. Beginning in 2003, ten years' worth of data will be displayed.

  EPA also improved the way SDWIS/FED data in Envirofacts are displayed. Major changes included combining violations and enforcement actions so that they are displayed in the same table (previously, users had to match violations and enforcement actions by looking at two different tables and matching the violations identification number), showing health-based violations separated from monitoring and other violations, and adding links to utilities' Consumer Confidence Reports (CCRs) on-line using EPA's new CCR catalog. Better descriptions of what violations and enforcement actions are, as well as additional links to state pages and contaminant fact sheets, were also added.
- Prioritized and corrected deficiencies already identified in the data entry process
- Accelerated the development and implementation of SDWIS/STATE
- Provided additional error check routines in SDWIS/FED
- Improved existing data entry tools such as the data entry troubleshooter's guide
- Accelerated efforts to develop new tools to simplify data retrieval, and accelerated efforts to improve existing reporting tools
- Developed an interim mechanism to enable utilities to confirm their data before they are officially accepted in SDWIS/FED

EPA Regions took additional steps to ensure that quarterly submissions are reviewed and errors are checked prior to the quarterly freeze in SDWIS/FED.

EPA and States drafted quality assurance manuals to help states and regions operate the drinking water program and report drinking water information.

3.2 Actions taken resulting from September 1999 Stakeholder Workgroup recommendations

The Stakeholder Workgroup reviewed the preliminary findings of the analyses used to quantify and characterize SDWIS/FED data quality in September 1999. Many of the actions taken or scheduled listed below resulted from their prioritized recommendations, which are listed in Appendix A.

3.2.1 EPA HQ actions taken

- Training:

  EPA staff have designed implementation and data reporting training courses for the Lead and Copper Rule Minor Revisions (LCRMR) and the Public Notification Rule (PN). Several courses have been conducted for states and regions.

  EPA has established a contractual arrangement for states and regions to obtain one-on-one, on-site data management assistance.

  EPA has expanded its offering of generic data entry and troubleshooting (i.e., correcting errors) training courses.

- The SDWIS/FED Edit/Update Summary Report has been completely redesigned to fully account for and document the processing results of each data submission file.

3.3 Future actions planned

3.3.1 EPA HQ Actions

- Provide additional training by:

  Developing a schedule for implementation and reporting training courses for the Chemicals/Radionuclides rules, the Surface Water Treatment Rule, the Total Coliform Rule, and developing training courses and materials for each new rule. The training will include implementation, compliance determination and reporting requirements.

- Improve the data verifications audits by:

  Revising the Data Verification Protocol to incorporate workgroup recommendations, completing a version of the Data Verification Protocol for states to use in conducting a self-audit, and completing 11 data verification audits in FY2000 (more if funds allow). If 17 audits were conducted per year, the data quality in each state could be assessed every 3 years (audits cost roughly $25,000 each).

- Complete a version of the error report which managers can use to help them improve data entry.
- Target attention to some states and regions, based on the results of individual state analyses and ongoing data verification audits. EPA will conduct meetings to address issues, target technical assistance and develop plans of action with such states and regions.
- Continue to calculate SDWIS/FED data quality including:

  National estimates for SDWIS/FED data quality at least every 3 years or more frequently if data from a sufficient number of data verifications analyses are available; ACR vs. SDWIS/FED analysis, national estimates of the timeliness of data reporting violations, and the number of states reporting violations by contaminant/rule and water system type annually; and error rates by error code quarterly.

3.3.2 EPA Regional Actions

- Conduct the errors analyses quarterly to determine which error conditions are occurring most frequently.

3.3.3 State Actions

States may take the following actions to improve data quality, but specific actions in each state will be contingent on its particular situation.

- Notify utilities annually of compliance monitoring schedules
- Implement and participate through Association of State Drinking Water Administrators (ASDWA) in peer reviews among states
- Conduct self-audits using the revised data verifications protocol
- Share software, tracking systems, and compliance determination modules among states that support rule implementation
- Evaluate current information management systems and consider adopting SDWIS/STATE
- Participate in EPA-provided training for rule implementation, reporting requirements, and data entry
- Develop and implement a quality assurance program

3.3.4 Joint EPA-State Actions

- Work together to establish goals for improving SDWIS/FED data quality at the national level, as assessed through data verifications results.
- Continue early involvement of states and regions in rulemaking with a focus on (1) streamlining reporting requirements and (2) simplifying rules to ease interpretation and implementation, including reporting requirements.

Potential categories for SDWIS/FED data quality goals:

- Overall inventory
- Overall enforcement actions
- Violations:
  TCR MCL   Other MCL
  SWTR TT   LCR TT
  M/R

3.4 Implementation process

The ASDWA/EPA Data Management Steering Committee (DMSC), in conjunction with the Data Sharing/Data Quality Committee (DSC), will continue to focus on data quality improvement issues identified in this report, and will propose future corrective actions and strategies for EPA and States.

Individual state-specific recommendations will be communicated to the states and EPA Regions through State Summary reports. Joint discussions will be conducted and an implementation schedule developed. Follow-up activities will be conducted through the normal mid-year and end-of-year program evaluation process. Generic state corrective actions will be pursued through the State/EPA annual Workplan process.

Formal implementation could begin as early as FY 2001. Many states have already begun state-specific corrective actions, as has EPA. Once finalized, appropriate standard operating procedures will be developed and incorporated in the EPA PWSS Data Management Quality Assurance Manual.

Collectively, steps already taken by EPA and States, and those planned, are expected to significantly improve the quality of data in SDWIS/FED. These steps should also improve public understanding of the high quality of drinking water supplied to consumers by most water systems in the United States.

Text Box 2

Improving the 4 Elements of SDWIS/FED Data Quality

There are 4 major elements of data quality**:**

1. **Completeness**—what percent of data that should be in SDWIS/FED is there?
2. **Accuracy**—how accurate are the data in SDWIS/FED?
3. **Timeliness**—what percent of violations data are being reported within a quarter after their compliance period end dates?
4. **Consistency**—are the regulations being interpreted consistently?

Actions taken or planned should improve these elements of SDWIS/FED data quality as follows:

	Completeness	**Accuracy**	**Timeliness**	**Consistency**
Early actions taken by EPA HQ				
Improved the way data are presented in the EPA Envirofacts website		X		
Corrected deficiencies in the data entry process	X	X	X	
Accelerated the development and implementation of SDWIS/STATE	X	X	X	X
Provided additional error check routines in SDWIS/FED		X		
Improved existing data entry tools such as the data entry troubleshooter's guide	X	X	X	
Accelerated efforts to develop new tools to simplify data retrieval		X		
Developed interim mechanism to enable utilities to confirm their data before they are officially accepted in SDWIS/FED	X	X		
EPA HQ actions taken resulting from September 1999 Stakeholder Workgroup recommendations				
Designed implementation and data reporting training classes for the LCRMR and PN Rules	X	X	X	X
Established arrangement for states and regions to obtain one-on-one, on-site data management assistance	X	X	X	
Expanded offering of generic data entry and troubleshooting training courses	X	X	X	
Redesigned SDWIS/FED Edit/Update Summary Report	X	X		
EPA HQ actions planned				
Provide additional rule-specific training for existing and upcoming rules, including implementation, compliance determination and reporting requirements.	X	X	X	X
Improve the data verifications audits to enable states to conduct self-audits; perform additional audits	X	X	X	X
Complete a version of the error report which managers can use to help them improve data entry	X	X	X	
Target poorer-performing states and regions; conduct meetings to discuss issues, target technical assistance and develop plans of action with such states and regions	X	X	X	X
Continue to quantify SDWIS/FED data quality	Benchmark DQ			
EPA Regional actions planned				
Conduct the errors analysis quarterly to determine which data entry error conditions are occurring most frequently	X	X	X	
State actions planned				
Notify utilities annually of compliance monitoring schedules	Reduce M/R viols.			
Implement and participate through ASDWA in peer reviews among states	X	X	X	X
Conduct self-audits using the revised data verification protocol	X	X	X	X
Share software, tracking systems, and compliance determination modules among states that support rule implementation	X	X	X	X
Evaluate current information systems and consider adopting SDWIS/STATE	X	X	X	X
Participate in EPA-provided training for rule implementation, reporting requirements, and data entry	X	X	X	X
Develop and implement a quality assurance program	X	X	X	X
Joint EPA-State actions planned				
Work together to establish goals for improving SDWIS/FED data quality, for specific categories of data, at the national level	X	X	X	X
Continue early involvement of states and regions in rulemaking with a focus on (1) streamlining reporting requirements and (2) simplifying rules to ease interpretation and implementation, including reporting requirements	X	X	X	X

PART II: DETAILED FINDINGS

4 National estimates of the quality of SDWIS/FED data

Part II provides details of the analyses conducted and the estimates of SDWIS/FED data quality. This section provides a definition of data quality, describes the analytical methodology used, and provides detailed estimates of the quality of inventory, violations and enforcement actions data in SDWIS/FED.

4.1 Data quality defined

Two questions need to be answered in order to estimate the quality of SDWIS/FED data:

1. **What should be in SDWIS/FED (and is missing)?**
2. **How accurate is what is in SDWIS/FED?**

There are four major elements of data quality. The first two are essentially a variation on the two questions above:

- **Completeness—**what percent of data that should be in SDWIS/FED is there?
- **Accuracy—**how accurate are the data in SDWIS/FED?

There are two additional elements of data quality:

- **Timeliness—**what percent of violations data are being reported within a quarter after their compliance period end dates? Timeliness is a component of Completeness
- **Consistency—**are the regulations being interpreted consistently?

4.2 Methodology

4.2.1 How EPA quantified data quality

This quantification is based on discrepancy rates for inventory, violations and enforcement action data. **Discrepancy rates** are defined as the percent of data that should be in SDWIS/FED that have errors, are missing, or that do not match between state databases and SDWIS/FED.

Overall data quality (for inventory, violations and enforcement actions data) is defined as the percent of data with no discrepancies. If, for example, 20% of the data have discrepancies, the SDWIS/FED data quality is 80%.

For violations and enforcement actions data, overall data quality can also be defined as the multiple of **Completeness** and **Accuracy**. Because inventory data are not exceptions-based data they are not quantified the same way. Instead, they are quantified as a single number. Accuracy is conditional on Completeness: it measures the accuracy of the data, given the data are complete.

For example, if there are:

100 violations that should be in SDWIS/FED,

and 60 make it in (Completeness)
and, of those, 48 are accurate (Accuracy),
Overall quality=60/100*48/60=48%

Timeliness is a component of **Completeness** and is included in the Completeness calculations; it was quantified separately in the frozen database analysis. **Consistency** is not quantified in this analysis, but is implicit, to some degree, in the data verifications.

4.2.2 Which estimates for which data

Next to each data type (in bold) is a list of parameters for which EPA calculated data quality estimates; below each are the analyses used to generate these estimates.

SDWIS/FED data quality estimates

Inventory—core data elements:

1. status (i.e., water system is active or inactive), 2. type of public water system (e.g., community, transient), 3. primary source of water, 4. population served, 5. number of service connections, 6. address, 7. name, 8. PWS ID

Overall SDWIS/FED data quality:

- Data verifications analysis
- Industry surveys

Violations—all violations

Overall SDWIS/FED data quality:

- Data verifications analysis

Completeness:

- Data verifications analysis

Accuracy:

- Data verifications analysis (with input from the Annual Compliance Report vs. SDWIS/FED analysis)
- Industry surveys

Timeliness:

- Frozen database comparison

Enforcement actions—all required to be reported to SDWIS/FED

Overall SDWIS/FED data quality:

- Data verifications analysis

Completeness:

- Data verifications analysis

Accuracy:

- Data verifications analysis
- Industry surveys

4.3 Perspective/context

These results in no way should be interpreted as a reflection of drinking water quality, which overall remains high. Nor does it question the accuracy of the data submitted by laboratories or water systems to states (inaccurate lab results, fraud, data falsification, etc.). The thousands of compliance decisions that are made correctly by state drinking water administrators are not enumerated. Only the violations and enforcement actions appear, because SDWIS/FED is an exceptions database (in other words, states do not provide sample data on regulated contaminants; they only report to SDWIS/FED when an "exception," such as a violation or enforcement action, has occurred). As will be shown, only a small percentage of systems have any health-based violations. Many states have taken corrective steps to improve their SDWIS/FED data quality since these data were gathered.

4.4 Data verifications

4.4.1 Background

The data verifications analysis is the only analysis that assesses the first key component of data quality for violations and enforcement actions data: **Completeness**, or the percentages of these data which should be in SDWIS/FED that are. The data verifications analysis also yields overall SDWIS/FED data quality estimates for inventory data (as do the inventory surveys).

The purpose of data verification audits is to determine whether a state is in compliance with that state's primacy agreement (since late 1996, auditors have considered guidance from Regions in addition to Federal regulations). Recommendations contained in the audit are intended to assist states in correcting deficiencies in their program and improve SDWIS/FED data quality.

An independent contractor has been performing data verifications since 1991. The contractor selects a (semi-) random sample of each type of water system in the state. During an audit, auditors primarily look at state files and database(s). The results are intended to be representative of the quality of drinking water data throughout the state with at least an 80% confidence level, and a 7.5% margin of error.

States have the opportunity to review the draft report and provide appropriate documentation required to adjust or revise the final report. Most states have accepted the final results of their data verification audits.

Prior to this analysis, data verification reports tabulated the number of systems having discrepancies. For this analysis, EPA tasked the contractor to re-tabulate the data on a data point basis—as a true SDWIS/FED data audit. That is, they compared data that should have been reported to SDWIS/FED to those that actually were reported, and cited reasons for each discrepancy. Now all data verifications are tabulated in this way.

For this analysis, EPA selected all data verifications done between 1996 and 1998. This included 29 data verification audits from 27 states. A total of 1,857 systems were audited. Some of the data verifications focused only on specific rules/contaminants. Results from the portion of the audit associated with the Lead and Copper Rule (LCR) are not included in this analysis due to questions of regulatory interpretation, which have not yet been resolved.

4.4.1.1 States included in this analysis, by EPA Region:

I	II	III	IV	V	VI	VII	VIII	IX	X
CT	VI	DE	AL	MI	LA	IA	SD	AZ	WA
MA		MD	FL	MN	NM	NE	WY		
ME		PA*	GA		OK				
NH		WV	NC		TX				
RI*									
VT									

* 2 audits were perfomed

4.4.1.2 Period of review for states reviewed during 1996-1998

Total Coliform Rule (TCR)	Most recent four quarters in SDWIS/FED
Nitrates	Most recent three calendar years
Nitrites	1993-1995
IOCs	1993-1995; back to 1990 if grandfathered
VOCs	1993-1996; back to 1988 if grandfathered
SOCs	1993-1995; back to 1990 if grandfathered
Radionuclides	Most recent two samples
Total Trihalomethanes	Most recent four quarters available in SDWIS/FED
Surface Water Treatment Rule	Most recent four quarters available in SDWIS/FED
Enforcement	Time period applicable to related violation

4.4.1.3 Summary of some states’ concerns about using data verifications results to quantify SDWIS/FED data quality

After EPA calculated SDWIS/FED data quality based on the data verifications, it shared the draft results with the states. Many states accepted the findings, and the methods used to derive them.

- One of the states’ most widespread concerns was that the public would misconstrue the quality estimates as an indication of how well states are running their drinking water programs, or as a measure of their drinking water quality. They felt a more accurate picture of state data quality would consider all the decisions a state is required to make, not just violation decisions. For example, a state may determine that a utility monitored correctly in eight out of ten instances. However, the state failed to issue one of the two violations which should have been issued. States noted that data quality in this case was really 90% (eight out of eight appropriate monitoring, one out of two failure-to-monitor instances results in a violation). In this analysis, only violation opportunities are considered. Since there were two violation opportunities and one of them was missed, this analysis would calculate SDWIS/FED data quality in this instance as 50%.
- A number of states pointed out that one improper determination could turn into multiple deficiencies. For example, if a system, due to being mis-categorized as a smaller system, collects 1 coliform sample per month instead of the 2 required, the data verifications will list a dozen violation discrepancies for the year. However, EPA should note that this is at least partially balanced by the fact that 1 missing sample is counted as 1 M/R discrepancy. Some sample bottles are to be used for several

contaminants, which, if missed, would result in up to 30 M/R violations (for a missing Synthetic Organic Chemicals (SOC) sample).

- A few data verifications were targeted to states having known data quality concerns. Some state reports are therefore better characterized as the "worst case" scenario.
- Some Federal requirements had just become effective in the time frame covered by the audits and many states were still in the process of adopting state rules and developing state data systems. Some of the data discrepancies are a function of normal and expected "start-up" problems. Some states felt that a snapshot taken today is likely to show a much better picture than one taken 3 years ago because many states have made data quality improvements since, and resulting from, their audits. Data verifications conducted in 1999 and after no longer review the previous compliance periods and therefore will be compared to the results in this analysis to measure the improvements suggested here.
- A few states contest their initial data verification audits. Some states believe that the data verification review team overlooked existing data (particularly monitoring results) and incorrectly determined that a violation had occurred when it had not. A number of states have pointed out errors in the data verifications findings which have since been investigated and corrected, and are reflected in this report.

Despite these concerns, EPA believes the findings are representative of SDWIS/FED data quality at the national level. Even slight biases (some of which tend to cancel each other out) do not significantly change the overall findings.

4.4.2 Confidence in findings

This is not a scientific survey and therefore statistical confidence intervals are not included for most of the point estimates. However, EPA is confident that the findings represent the quality of SDWIS/FED data at the national level.

First, the data verifications audits are designed to be representative of the quality of drinking water data throughout the state with at least an 80% confidence level and a 7.5% margin of error. In addition, the audits have undergone scrutiny: in the summer of 1999, states and regions had an opportunity to review the findings of their audits, and any errors found were corrected.

Second, EPA considers the summation of the 29 audits in 27 states to be representative of the quality of drinking water data at the national level. This was ascertained after EPA modeled the individual state findings mathematically using Bayesian statistics; the resulting probability curve was found to have a normal distribution.

Third, EPA looked at data quality from many perspectives, and has compared estimates with the results of other analyses wherever possible. As will be discussed, findings from other analyses corroborated the data verifications findings.

4.4.3 State Annual Compliance Report (ACR) vs. SDWIS/FED

The data verifications analysis has a category for violations discrepancies between state databases and SDWIS/FED. However, it does not indicate what portion of these discrepancies represent under-reporting (data which is in state databases but not SDWIS/FED) and over-reporting (data which is in SDWIS/FED but not state databases). It is necessary to make this distinction in order to yield estimates of Completeness and Accuracy.

To accomplish this, EPA compared calendar year 1997 ACR data reported using state databases to 1997 data in SDWIS/FED. EPA calculated ratios of the magnitude of under-reporting to over-reporting for Chemical, Total Coliform Rule (TCR), and Surface Water Treatment Rule (SWTR) health-based violations and monitoring/reporting violations. These ratios were input to the data verifications analysis to enable EPA to calculate estimates for Completeness and Accuracy for violations data.

4.4.4 Inventory data

4.4.4.1 Estimates by parameter, and overall

Four percent of 8 required inventory data points had discrepancies, or errors. In other words, the overall SDWIS/FED inventory data quality is estimated to be 96%, as shown below.

	Status: Active/Inact.	Water system Type	Primary source	Population	# Service connections	Address	Name	PWS ID	Overall
Number of Systems Reviewed	2,032	2,014	1,997	1,996	1,996	1,979	1,996	1,996	16,006
Discrepancies:									
Number	58	61	39	184	161	99	41	3	646
Percent	2.9%	3.0%	2.0%	9.2%	8.1%	5.0%	2.1%	0.2%	4.0%
SDWIS/FED data quality	**97%**	**97%**	**98%**	**91%**	**92%**	**95%**	**98%**	**100%**	**96%**

Each water system has 1 chance for a discrepancy for each parameter reviewed. The “Overall quality” column uses the sum of water systems reviewed for each parameter, which represents the total opportunities for a discrepancy.

The population served and # service connections parameters had the most discrepancies. A discrepancy in either of these categories is counted as such if the difference is greater than 10%. Under several drinking water rules, the number of samples required to be taken is based on the population served and therefore its accuracy is important.

4.4.4.2 Reasons for discrepancies

About one-half of the discrepancies were due to file inconsistencies between data in state files and the state database(s); another one-third were due to inconsistencies between data in state database(s) and SDWIS/FED; most of the remaining one-sixth were due to late reporting, or no data found in state files.

4.4.4.3 Estimates by system type and size

Results from the data verifications analysis were very similar across system types, as shown below. None of the quality estimates for the 8 parameters listed above differed by more than 4%.

CWS	97%
NTNCWS	96%
TNCWS	95%

Unfortunately, the results of the data verifications analysis cannot be categorized by system size. The only way to get any approximation using this data is to look at system types as a proxy for system size. The information below lists the average population served by system type (from the 98Q4 frozen database, which was frozen in January 1999).

CWS	4,645
NTNCWS	308
TNCWS	175

The average system size for NTNCWSs and TNCWSs is in the Very Small size category (25-500 population served), and for CWS the Medium size category (3,301-10,000). If these results can serve as a proxy for system size, then it appears that data quality may be similar across size categories. The industry surveys, discussed later, provide a direct measure of SDWIS/FED inventory data quality by system size so this report addresses this issue in Section 4.5.3.3.

4.4.5 Violations data

4.4.5.1 Estimates by violation type

Listed below are SDWIS/FED data quality estimates for violations data. The first line of the table shows the percent of systems (by violation type) having any violations.

Less than 10.4% of all systems audited in the data verifications had any Maximum Contaminant Level (MCL) violation, and less than 10% of the surface water systems audited had Surface Water Treatment Rule (SWTR) Treatment Technique (TT) violations. The estimate that slightly less than 78% of systems had M/R violations is based on the finding that 78% of all systems audited had at least one violation of any type, and M/R violations account for 94% of all violations. The 78% estimate also includes the small number of systems which only had LCR violations (earlier versions of the analysis included estimates for LCR, and it was not possible to subsequently remove LCR from this statistic).

These percentages of systems having violations lend a systems perspective. They are not part of the calculations of SDWIS/FED data quality, which is based on a data point perspective. The remainder of the table reflects a data point perspective.

	TCR MCL	Total Other MCL	Total MCL	SWTR TT	Total M/R
% systems w/ violations	6.1%	< 4.3%	<10.4%	9.6%	<78%
Number of Violations	162	59	221	94	5,090
Discrepancies					
Number	52	50	102	87	4,613
Percent	32%	85%	46%	93%	91%
% Completeness	68%	19%	55%	11%	10%
% Accuracy	99%	79%	97%	67%	95%
SDWIS/FED data quality	**68%**	**15%**	**54%**	**7%**	**9%**

Legend:

- TCR: Total Coliform Rule, applicable to all water systems. Coliforms pose an acute health risk
- MCL: Maximum Contaminant Level violation
- TT: Treatment Technique violation (MCLs and TTs are health-based violations)
- M/R: Monitoring/Reporting violation

TCR MCL data will serve as an example to describe this table:

- 6.1% of systems reviewed incurred or should have incurred a total of 162 MCL violations
- Of the 162 violations, there were 52 discrepancies, or errors. The discrepancy rate is 32%, and the corresponding SDWIS/FED data quality estimate is 68% (100%-32%).
- Completeness and Accuracy—68% of the violations that should be reported in SDWIS/FED made it in, and of the violations in SDWIS/FED, 99% are accurate.

According to these estimates, roughly 2/3 (68%) of all TCR MCL violations were reported completely and accurately. The SDWIS/FED data quality is 15% for Other MCLs, 7% for SWTR TTs, and 9% for M/R violations.

Overall, the data that do make it into SDWIS/FED are accurate. In fact, 99% of the TCR MCL violations, 79% of Other MCL violations, and 95% of M/R violations listed in SDWIS/FED are accurate. However, only, 2/3 of SWTR TT violations listed in SDWIS/FED are estimated to be accurate. In other words, there may be some over-reporting of SWTR TTs in SDWIS/FED.

The weak link in data quality is the large number of violations that never make it to SDWIS/FED (as estimated by Completeness). Only 1 out of every 9 SWTR TT violations that should be in SDWIS/FED make it in (11% Completeness), and only 1 out of every 10 M/R violations make it in.

4.4.5.2 Reasons for discrepancies

The data verifications include several categories, or reasons, for violations discrepancies.

Reason	TCR MCL	Other MCL	Total MCL	SWTR TT	M/R
Not in state database					
No data found in state files	0	0	0	0	3,492
Insufficient samples	0	0	0	0	205
Different implementation policies	31	22	53	72	417
Other	4	0	4	0	56
In state database(s) but not SDWIS/FED	16	26	42	11	252
In SDWIS/FED but not state database(s)	1	2	3	4	25
Total	52	50	102	87	4,613

M/R violations discrepancies account for the majority (94%) of all the discrepancies, with the largest category being “no data found in state files.” This category applies to

M/R violations only. If, for example, required sample results could not be found in any state files, a discrepancy would be cited if the state did not issue a M/R violation. This could also occur if water systems were told they could reduce the monitoring frequency for some requirements, but no record of a waiver having been issued was found.

The "Different implementation policies" category means that the state did not determine compliance in accordance with their state primacy agreement. Since late 1996, auditors have also been factoring in any additional guidance provided by EPA Regional offices. Thus, as long as a state acts in accordance with its own EPA-approved regulations, or formal interpretive guidance issued by the region, no discrepancy is issued.

The last category listed, "In SDWIS/FED but not state database(s)," represents over-reporting. All the other categories represent under-reporting. Overall, 99.3% of all violation discrepancies found in the data verifications analysis are estimated to be from under-reporting. Only 32 out of the 4,802 violation discrepancies found (<0.7%) are estimated to be from over-reporting. These estimates are based in part on ratios of under- to over-reporting identified from the ACR analysis.

Most violations discrepancies are related to compliance determination at the state level, which consist of violations which never made it into state databases. The remaining discrepancies (i.e., those listed in the last two rows of the above table) are related to data flow between state files and SDWIS/FED. However, since monitoring/reporting discrepancies comprise 94% of the total number of discrepancies an overall number shouldn't represent this. A more precise picture is portrayed when the discrepancy categories are analyzed by violation type:

Breakdown of discrepancies	TCR MCL	Other MCL	Total MCL	SWTR TT	M/R
Compliance determination	67%	44%	56%	83%	94%
Data flow	33%	56%	44%	17%	6%

As shown in the table above, only one-third of all TCR MCL violation discrepancies occur between state files and SDWIS/FED (17/52). Over one-half of Other MCLs (28/50), one-sixth of SWTR TTs (45/102), and only 6% of all monitoring/reporting violation discrepancies (277/4,613) occur between state files and SDWIS/FED. Other analyses will look at some reasons for these data flow discrepancies, including the frozen database analysis, which looks at Timeliness (were some violations merely entered late?), and the errors analysis (were some violations rejected at data entry?).

4.4.5.3 Estimates by rule/contaminant

The data verifications also listed violations data by rule/contaminant. There were neither sufficient data points to calculate quality estimates for some of the Chemical MCLs, nor sufficient data points to calculate estimates of Completeness and Accuracy. Again, a systems perspective, listing the percentage of systems having any violations, precedes the SDWIS/FED data quality estimates.

	MCLs								TTs
	TCR	IOCs	Nitrate	Nitrite	SOCs	VOCs	TTHMs	Rads	SWTR
# systems reviewed	1,857	1,025	1,489	1,489	1,025	1,026	83	523	395
# systems w/ violations	113	10	19	3	0	2	1	2	38
% systems w/ violations	6.1%	1.0%	1.3%	0.2%	0.0%	0.2%	1.2%	0.4%	9.6%
# Violations	162	12	37	4	0	2	1	3	94
# Discrepancies	52	11	32	3	0	1	0	3	87
% Discrepancies	32%	92%	86%	75%	0%	50%	0%	100%	94.7%
SDWIS/FED data quality	**68%**	**8%**	**14%**	**25%**	*****	*****	*****	*****	**5%**

* insufficient data

	M/Rs								
	TCR	IOCs	Nitrate	Nitrite	SOCs	VOCs	TTHMs	Rads	SWTR
# systems reviewed	1,857	1,025	1,489	1,489	1,025	1,026	83	523	395
# systems w/ violations	480	175	507	224	263	257	6	111	83
% systems w/ violations	26%	17%	34%	15%	26%	25%	7%	21%	21%
# Violations	1,289	193	964	235	877	722	12	163	635
# Discrepancies	1,034	174	844	210	864	686	12	161	628
% Discrepancies	80%	90%	88%	89%	99%	95%	100%	99%	99%
SDWIS/FED data quality	**20%**	**10%**	**12%**	**11%**	**1%**	**5%**	**0%**	**1%**	**1%**

Legend:
MCL: Maximum Contaminant Level violation
TT: Treatment Technique violation
(MCLs and TTs are health-based violations)
M/R: Monitoring/Reporting violation

TCR: Total Coliform Rule
IOC: Inorganic Chemicals
SOCs: Synthetic Organic Chemicals
VOCs: Volatile Organic Chemicals
TTHMs: Total Trihalomethanes
Rads: Radionuclides
SWTR: Surface Water Treatment Rule

SDWIS/FED data quality for TCR data is significantly higher than for other rules or contaminants. The data quality estimates for SOCs, TTHMs, Rads, and SWTR averaged 1% quality, or less. The vast majority of these discrepancies were due to under-reporting—specifically, no data found in state files. In other words, no more than 1 of every 100 SOC, TTHM, Rad, and SWTR M/R violations are reported to SDWIS/FED.

4.4.5.4 Estimates by system type

These data by water system type were similar enough to be counted together, as shown below. By doing so, the accuracy of our estimates increased significantly, since more data points yield better estimates.

	TCR MCL	SWTR TT	M/R
CWS	69%	9%	9%
NTNCWS	67%	11%	7%
TNCWS	68%	0%	14%
Overall	68%	7%	9%

Again, there are not enough data points to calculate estimates for Completeness and Accuracy, nor are there sufficient data to estimate quality of Other MCLs by system type.

Unfortunately, the results of the data verifications analysis cannot be sorted by system size. The industry surveys can be categorized in this way, as will be discussed later.

4.4.6 Enforcement actions data

4.4.6.1 Estimates by system type, and overall

Estimates for formal SDWIS/FED enforcement actions data quality are preceded by a systems perspective.

	CWS	NTNCWS	TNCWS	Total
# Systems Reviewed	696	548	562	1,806
# Systems with Enforcement Actions	163	122	75	360
% Systems with Enforcement Actions	23%	22%	13%	20%
# Enforcement Actions	505	305	222	1,032
# Discrepancies—under-reporting	55	53	29	137
# Discrepancies—over-reporting	37	17	24	78
# Discrepancies—incorrect reporting	29	22	21	72
Total discrepancies	121	92	74	287
% Discrepancies	24%	30%	33%	28%
% Completeness	89%	83%	87%	87%
% Accuracy	85%	85%	77%	83%
SDWIS/FED data quality	**76%**	**70%**	**67%**	**72%**

Data in the "Total" column will serve as an example to describe this table:

- System perspective—of the 1,806 systems reviewed in the audits, 360, or 20%, had enforcement actions.
- Of the 1,032 enforcement actions listed for these 360 systems, there were 287 discrepancies. The discrepancy rate is 287/1,302 or 28%.
- Overall, 87% of the data that should be in SDWIS/FED make it in (Completeness), and 83% of the enforcement actions in SDWIS/FED are accurate.

The calculation for Completeness is based on the number of discrepancies that represent under-reporting. Here the data verifications are clear as to which actions were not reported to SDWIS/FED. The calculation for Accuracy is based on Over-reporting (missing from state files) as well as for Incorrect reporting (which occur if the dates listed in SDWIS/FED are off by more than a month).

The quality estimates are similar across system types: all were within 4% of the combined average.

4.5 Industry surveys

4.5.1 Background

Both the National Rural Water Association (NRWA), in conjunction with the Association of Drinking Water Administrators (ASDWA), and the American Water Works Association (AWWA) volunteered to survey their water systems.

The objective of this effort was to get data quality estimates from water systems directly. Indeed, this is the only analysis that goes upstream of state records. From this analysis EPA derived overall SDWIS/FED data quality estimates for inventory data, and Accuracy estimates for violations and enforcement actions data. Operators were not asked to assess the Completeness of violations and enforcement actions data, but only the Accuracy of those listed in SDWIS/FED. Another objective of this effort was to provide states with feedback from this effort to help them investigate and correct potential errors that may exist. Any corrections they make will be reflected in SDWIS/FED in the next quarterly update after the state's corrections are submitted to SDWIS/FED.

4.5.2 Survey design

Water systems surveyed received a printout of their inventory, violations, and enforcement action data from SDWIS/FED. Water system operators were asked to indicate whether each data point was correct, incorrect, or to indicate "DK" if they did not know. Each data point marked "DK" was removed from the survey analysis so as not to artificially lower the discrepancy rates.

AWWA sent surveys to all water systems serving more than 10,000 people that incurred at least one violation between FY1993 and FY1997. Of the 2,222 surveys sent, 684 were completed and returned, resulting in a 31% response rate (25% is a typical response rate for mailed surveys).

NRWA/ASDWA surveyed active, current systems serving fewer than 10,000 people that incurred at least one violation between FY1993 and FY1997. A random sample of 40 CWSs and 5 NTNCWSs were selected for each state. Of 2,549 surveys sent, 439 were completed and returned from 23 states. The response rate was 17% overall, and 39% from the 23 states that participated.

As discussed below, in both surveys, water systems that did not respond had a higher average number of violations than those that did. The effect of this self-selection bias on the results of this analysis is unclear.

Two systems were removed from the AWWA survey. A system in New Jersey disputed all of their 751 violations. This may be a case of over-reporting, but the inclusion of this single system in the survey would have resulted in overall discrepancy rates four times higher. A system in PA with 718 violations was removed because it was not clear how to categorize their violations. On their survey sheets, the water system indicated "DK." In a letter they sent with their completed survey they did not dispute any of them, and in fact explained how several of them occurred. In a telephone interview, they disputed all of them.

An EPA contractor conducted telephone interviews with 7 water system operators to evaluate how they filled out the survey, and to investigate potentially "extreme" responses—water systems which disputed either all or none of their violations data.

The contractor found the presence of response bias in the violations and enforcement actions responses: a number of water systems contacted said they left violations and enforcement actions data points blank, rather than indicating "DK," if they were unsure whether the data points were correct or not. The magnitude of this bias is unclear.

4.5.3 Inventory data

Operators did a thorough job in evaluating their inventory data. Since each water system had 7 required data points to evaluate, the results from each water system are counted equally. This is in contrast to violations and enforcement actions data, where a few systems having hundreds of violations, for example, significantly increase the average violations discrepancy rates. The result shows a fairly high degree of confidence in these inventory estimates.

4.5.3.1 Estimates by required parameter, and overall

	Status: Active/Inact.	Water system Type	Primary source	Population	# Service connections	Address	Name	Overall
AWWA survey	100%	100%	90%	85%	86%	87%	97%	92%
NRWA survey	99%	98%	97%	84%	85%	87%	97%	93%
Data verifications	97%	97%	98%	91%	92%	95%	98%	96%

Public water system identification number (PWS ID) was not assessed in the surveys. A discrepancy in the Population served and # service connections is counted as such if the difference is greater than 10%, as was done in the data verifications analysis.

As shown above, the overall SDWIS/FED data quality estimates are very close to but slightly lower than the estimates from the data verifications analysis. The surveys estimated slightly lower SDWIS/FED data quality for primary source (AWWA survey only), population served, service connections, and address.

The surveys also asked water system operators to evaluate some optional data parameters. This information was requested to estimate the quality of the currently optional data that would become required as of January 2000.

4.5.3.2 Estimates by additional parameters

	Primary Contact	Phone	Owner category	County 1	County 2	Principal city	Principal county
AWWA survey	93%	65%	96%	97%	100%	56%	95%
NRWA survey	94%	70%	91%	91%	100%	76%	98%

4.5.3.3 Estimates by system type and size, for required data

	System type	
AWWA survey	CWS	92%
NRWA survey		92%
Data verifications		97%
NRWA survey	NTNCWS	97%
Data verifications		96%

	Size category	
NRWA survey	Very Small	93%
“	Small	92%
“	Medium	91%
AWWA survey	Large	92%
“	Very Large	91%

These estimates by system type are slightly lower than those estimated by the data verifications for CWSs, and they are very close for NTNCWSs.

As discussed above, EPA was not able to calculate SDWIS/FED data quality estimates by system size category in the data verifications analysis. Fortunately, EPA was able to do this in the industry surveys. As shown above, the results across size categories are very close and show high data quality for required inventory data selected for this analysis.

4.5.4 Violations data

As described above, the surveys yielded estimates of the Accuracy of the data in SDWIS/FED; they did not assess the Completeness of the data (the % of data that should be in SDWIS/FED that made it in). These Accuracy estimates are of uncertain value, due to the fact that some water system operators may have left data points blank because they were not sure whether or not a violation was correct (instead of indicating that they did not know). This dilutes the discrepancy rates to an unknown degree.

In addition, there may be some non-response bias: water systems included in the survey that did not respond averaged 40% more violations in the AWWA survey and 58% more violations in the NRWA survey than those that did. The effect of this bias is unclear.

Ninety-six percent (96%) of systems in the AWWA survey, and 91% in the NRWA survey, did not dispute any of their violations. Overall, these Accuracy estimates are very close to those from the data verifications analysis.

4.5.4.1 Accuracy estimates by violation type

	Total MCL	SWTR TT	Total M/R
NRWA survey	96%	91%	97%
AWWA survey	99%	99%	96%
Data verifications	97%	67%	95%

The Accuracy estimates for Total MCLs and Total M/Rs are very similar to those from the data verifications analysis. However, the surveys estimated a higher Accuracy of SWTR TTs than did the data verifications.

4.5.4.2 Accuracy estimates by rule/contaminant

	MCLs								**TTs**	
	TCR	IOCs	Nitrate	Nitrite	SOCs	VOCs	TTHMs	Rads	SWTR	LCR
NRWA survey	97.0%	n/a	89.9%	*	*	100%	*	100%	90.6%	100%
AWWA survey	99.4%	100%	100%	*	100%	100%	97.0%	100%	99.4%	96.8%

* insufficient data

	M/Rs									
	TCR	IOCs	Nitrate	Nitrite	SOCs	VOCs	TTHMs	Rads	SWTR	LCR
NRWA survey	93.6%	100%	94.2%	100%	100%	99.7%	*	100%	100%	88.1%
AWWA survey	91.7%	99.2%	96.8%	100%	76.7%	98.4%	94.0%	100%	69.3%	100%

* insufficient data

Overall, the surveys estimate very high Accuracy of these contaminants/rules. In other words, water system operators disputed very few of the violations listed in SDWIS/FED.

Unfortunately, EPA was not able to calculate comparable Accuracy estimates for violations data from the data verifications analysis. There were an insufficient number of data points available at the rule/contaminant level. Therefore, a direct comparison of the results listed above to the data verifications analysis is not possible.

4.5.4.3 Accuracy estimates by system type and size

	System type	Accuracy—all violations
AWWA survey	CWS	95%
NRWA survey	CWS	97%
”	NTNCWS	95%

	Size category	Accuracy—all violations
NRWA survey	Very Small	97%
“	Small	96%
“	Medium	95%
AWWA survey	Large	96%
“	Very Large	90%

Overall Accuracy estimates are very close across system types. By system size they are very close as well, with the exception that Very Large systems are 5-7 percentage points lower.

4.5.5 Enforcement actions data

Again, these Accuracy estimates are of uncertain value, due to the fact that some water system operators may have left data points blank instead of indicating “DK” on their surveys. This dilutes the discrepancy rates to an unknown degree.

4.5.5.1 Accuracy estimates by system type and size

	System type	Accuracy
AWWA survey	CWS	98%
NRWA survey		99.6%
Data verifications		89%
NRWA survey	NTNCWS	99%
Data verifications		83%

	Size category	Accuracy
NRWA survey	Very Small	99%
“	Small	99.8%
“	Medium	99%
AWWA survey	Large	99%
“	Very Large	98%

The Accuracy estimates are higher than those estimated in the data verifications analysis. In addition, the survey findings indicate that the Accuracy is very similar across system types and sizes.

4.6 Frozen database comparison—Timeliness estimates

A violation is due to be reported to SDWIS/FED within 90 days after its compliance period end date. This analysis quantifies how long it has taken for FY1997 violations to be reported.

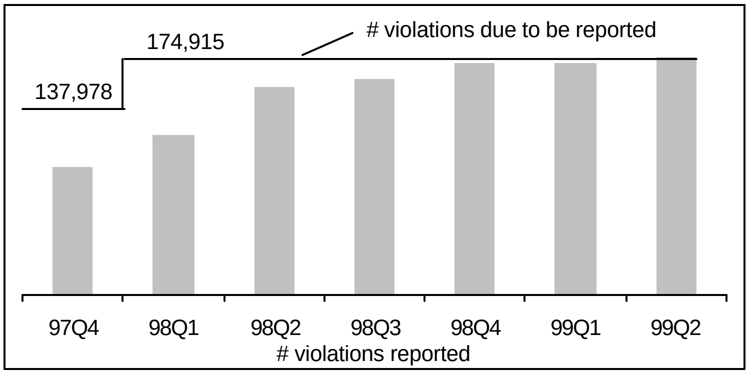
SDWIS/FED databases have been “frozen” quarterly since 1997. These frozen databases enable EPA to look at what data were in SDWIS/FED during set time periods. This analysis compares fiscal year 1997 data reported in each of the seven quarterly databases frozen since January 1998.

The following estimates are based on violations reported to SDWIS/FED by July 1999. This analysis assumes that all FY1997 violations which were going to be reported actually were reported by July 1999 (7 quarters after all violations were due). Data for North Carolina is not included in these estimates. Their reporting of violations data was highly erratic, which skewed the results.

There were 137,978 violations in FY1997 with end dates at or before September 30, 1997, which were due to be reported by December 31, 1997. Similarly, there were an additional 36,937 violations with end dates between October 1 and December 31, 1997, and these were due to be reported by March 31, 1998. Therefore, a total of 174,915 FY1997 violations were due to be reported not later than March 31, 1998.

There were also 12,849 violations having end dates later than December 31, 1997; they are not included in this analysis. Most had significantly later end dates and would not be due to be reported before July 1999.

Timeliness	97Q4, frozen Jan '98	98Q1, frozen Apr '98	98Q2, frozen Jul '98	98Q3, frozen Oct '98	98Q4, frozen Jan '99	99Q1, frozen Apr '99	99Q2, frozen Jul '99
# violations reported	94,484	118,318	153,988	158,752	170,793	170,647	174,915
# that should have been reported	137,978	174,915	174,915	174,915	174,915	174,915	174,915
% reported by each frozen database	**68%**	**68%**	**88%**	**91%**	**98%**	**98%**	**100%**



At the national level, this analysis indicates that 68% of FY1997 violations that should be in SDWIS/FED by December 31, 1997 made it in on time, and that 68% of all the violations that should have been reported by March 31, 1998 actually were reported by then. Late reporting is a component of Completeness. As can be seen above, late reporting is a significant problem.

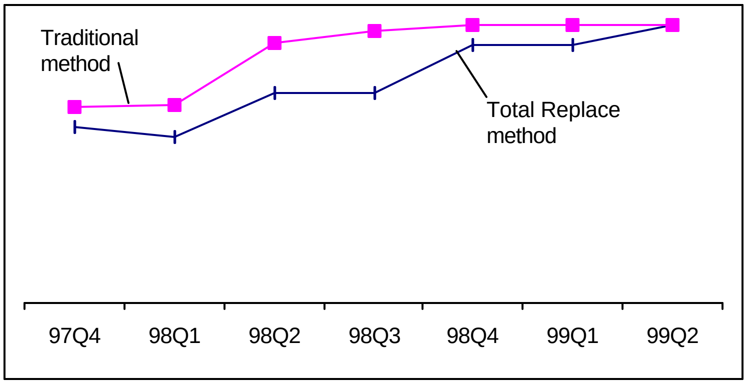
It was not possible to factor Timeliness estimates into the SDWIS/FED data quality estimates since the period of review for most contaminants/rules in the data verifications audits was primarily 1993-1998—most of which occurred before late 1997 when EPA began to "freeze" SDWIS/FED databases.

EPA was able to categorize Timeliness using the two methods of data entry to SDWIS/FED. EPA used data from the errors analysis to determine which state used which data entry method.

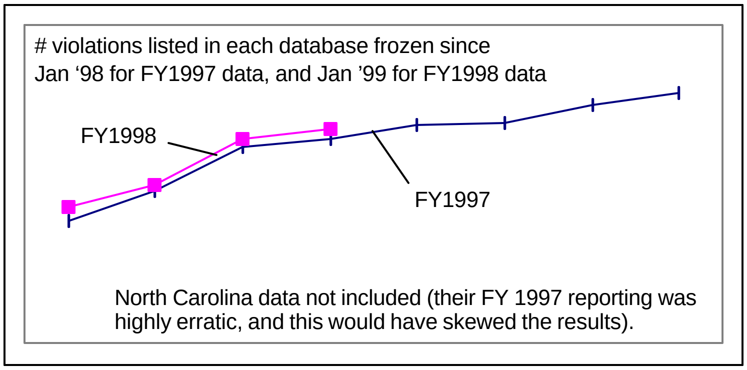
- Traditional method, wherein only new, modified, or deleted information is transmitted
- Total Replace method wherein the state sends a complete data set every quarter and totally over-writes all data previously submitted.

Violations appear to be reported in a more timely manner when the Traditional method is used to report violations, compared to the Total Replace method:

% reported by each frozen database	97Q4, frozen Jan ‘98	98Q1, frozen Apr ‘98	98Q2, frozen Jul ‘98	98Q3, frozen Oct ‘98	98Q4, frozen Jan ‘99	99Q1, frozen Apr ‘99	99Q2, frozen Jul ‘99
Total Replace method	63%	60%	75%	75%	92%	93%	100%
Traditional method	71%	71%	94%	98%	100%	100%	100%



As can be seen, many states have been adding and modifying FY1997 violations data several quarters after they were due. Through 1996, there does not appear to have been nearly as much volatility in the data. This may be the result of attention focused on correcting discrepancies between State Annual Compliance Reports and SDWIS/FED which were first identified in the 1996 and 1997 reports. An almost identical trend is occurring with FY1998 data, as illustrated below:


# violations listed in each database frozen since
Jan ‘98 for FY1997 data, and Jan ’99 for FY1998 data

North Carolina data not included (their FY 1997 reporting was highly erratic, and this would have skewed the results).

4.7 Comparison of SDWIS/FED to Envirofacts

The public sees SDWIS/FED data as displayed in Envirofacts, EPA’s multimedia website. One aspect of this analysis was to compare data in SDWIS/FED to Envirofacts to ensure that no errors are introduced in transfer of data from SDWIS/FED to Envirofacts. All data from 250 water systems selected at random were compared in the two databases to identify any data transfer errors. No errors were found.

5 Additional data quality analyses

5.1 States' reporting of violations data

As part of a further analysis of under-reporting identified initially in the data verifications analysis, EPA looked at the Annual Compliance Report (ACR) comparison to SDWIS/FED. The ACR vs. SDWIS/FED analysis indicated that several states did not report any violations at all in CY1997 for certain contaminants/rules. Some of the non-reporting was attributable to late reporting. Some states that did not report in 1997 had reported in other years.

To factor out late reporting, and to get a more comprehensive picture of non-reporting of certain violations by state, EPA queried the SDWIS/FED database frozen in October 1999 and listed all violations reported by each state between FY1993 and FY1998. It found that over a dozen states have never reported chemical rule violations for any NTNCWSs or TNCWSs, and half have never reported Radiological rule violations for CWSs.

Clearly, some of the non-reporting is attributable to states simply not having any violations to report. However, in light of the magnitude of under-reporting estimated in the data verifications analysis, and given the percentages of systems estimated to have violations, by rule, many of these "blanks" represent a problem. These "blanks" are being evaluated in state-by-state summaries of SDWIS/FED data quality.

The two tables below only include situations where the state has certain systems subject to a rule. One state has no NTNCWSs (Alaska), and the SWTR has no impact in states without any surface water systems in a system type category. One state/territory has no surface water CWSs, 7 have no NTNCWSs, and 13 have no surface water TNCWSs.

5.1.1 Number of the 52 states/territories that have never reported any violations, by rule, between FY1993 and FY1998

Below is a list of states/territories that have never reported a violation in this six-year period.

	TCR		Chemicals		RADs		LCR		SWTR	
	MCL	M/R	MCL	M/R	MCL	M/R	TT	M/R	TT	M/R
CWS	0	1	5	8	25	23	21	1	4	18
NTNCWS	0	0	13	14	Rule applies to CWS only		25	4	16	27
TNCWS	1	1	12	14			Does not apply		11	20

This can also be shown in percentages, which will facilitate a comparison with the next table.

	TCR		Chemicals		RADs		LCR		SWTR	
	MCL	M/R	MCL	M/R	MCL	M/R	TT	M/R	TT	M/R
CWS	0%	2%	10%	15%	48%	44%	40%	2%	8%	35%
NTNCWS	0%	0%	25%	27%	Rule applies to CWS only		49%	8%	36%	60%
TNCWS	2%	2%	23%	27%			Does not apply		26%	48%

5.1.2 Percent non-reporting of violations, by type, between FY1996 and FY1998

It is informative to look at the percentage of non-reporting, to count the percentage of "blanks" in each year. The table below lists the percent of non-reporting that occurred between FY1996 and FY1998 by contaminant/rule. There are 156 opportunities to report violations in each box below (52 states*3 years), less the number of states not counted, as described above.

	TCR		Chemicals		RADs		LCR		SWTR	
	MCL	M/R	MCL	M/R	MCL	M/R	TT	M/R	TT	M/R
CWS	0%	3%	22%	28%	62%	56%	55%	28%	17%	63%
NTNCWS	2%	5%	46%	38%	Rule applies to CWS only		70%	39%	55%	81%
TNCWS	5%	6%	40%	36%			Does not apply		53%	72%

This shows that almost all states have reported both TCR MCL and M/R violations in each year. The other rules have had significantly less reporting. For each rule and violation type, the most reporting has been done for CWSs. NTNCWSs and TNCWSs fared about the same as each other, but were reported less frequently than CWSs.

Comparing this table to the one above it shows that the percentage of "blanks" is in some cases significantly higher than when merely considering states that have never reported. In other words, the states that have reported for specific rules/contaminants have not done so in each year. For example, 10% of states have never reported a Chem MCL (which accounts for 10% of the "blanks"), but there are 22% "blanks" for Chem MCLs.

5.1.3 Percent non-reporting of violations, by year

The level of non-reporting in each year has been fairly steady, although it increased in 1998. EPA calculated the statistics below by dividing the total number of "blanks" each year by the total number of opportunities to report violations.

1998	64%
1997	59%
1996	58%
1995	58%
1994	53%
1993	58%

Again, some of these blanks represent states that simply had no violations in a category during a year. State-by-state summaries of SDWIS/FED data quality take a closer look at this issue.

5.2 Comparison of states' reporting of Annual Compliance Report (ACR) data to SDWIS/FED data

5.2.1 Background

This analysis highlighted differences between data in state databases and files and SDWIS/FED. These differences were analyzed numerically using the 1997 ACR data. States were also asked to identify reasons for discrepancies between what they reported for the 1996 and 1997 ACR and what is in SDWIS/FED.

5.2.2 Under- and over-reporting between state databases and SDWIS/FED

In this exercise, EPA calculated ratios of under- to over-reporting. These results were also used in the data verifications analysis.

The data verifications list discrepancies between state databases and SDWIS/FED, but do not divide them into over-reporting, under-reporting, and incorrect reporting (in the case of incorrect reporting, the violation exists in both databases but does not match). In order to calculate estimates for Completeness and Accuracy, EPA to ascribed discrepancies to either over-reporting or under-reporting (it is not possible to get numerical estimates of incorrect reporting). This will also enable EPA to compare accuracy estimates from the data verifications analysis to the industry surveys.

The ACR vs. SDWIS/FED analysis used 1997 ACR data. The ratios of under-reporting to over-reporting, by rule and overall, are shown below:

TCR		Chem		SWTR		LCR		Total			Overall
MCL	M/R	MCL	M/R	TT	M/R	TT	M/R	MCL	TT	M/R	Total
21.0	3.0	10.5	136.5	3.3	37.3	14.6	19.0	16.3	4.5	16.2	15.5

Significantly more under-reporting than over-reporting of violations was found. For example, of the 1997 ACR violations reported using state databases vs. SDWIS/FED, the magnitude of overall under-reporting was more than 15 times as great as the magnitude of over-reporting.

Here is how these estimates were calculated:

First, EPA excluded states that reported using SDWIS/FED, since EPA wants to compare what is in state databases to what is in SDWIS/FED. EPA also excluded Chemical M/R violations for one state that listed 21,807 violations in their state database and only 98 in SDWIS/FED; these numbers were an anomaly, and they skewed the overall results.

Next, instances of over-reporting and under-reporting were summed separately. For each, differences were taken (between the totals for violations in state databases and in SDWIS/FED).

Finally, the difference, or number of discrepancies, for under-reporting was divided by the difference for over-reporting.

5.2.3 Minimum discrepancy rates between state databases and SDWIS/FED

Along with the ratios calculated above, it is informative to look at the discrepancy rates between 1997 ACR data in state files and SDWIS/FED. These estimates are listed below:

TCR		Chem		SWTR		LCR		Total			Overall
MCL	M/R	MCL	M/R	TT	M/R	TT	M/R	MCL	TT	M/R	Total
15%	31%	40%	41%	20%	38%	86%	68%	18%	26%	39%	37%

These discrepancy rate estimates are minimum estimates since in order to generate them EPA had to assume that all violations match between state databases and SDWIS/FED. For example, if there are 6 violations in a state's database and 10 in SDWIS/FED, we've assumed that these 6 match, resulting in 4 instances of over-reporting. The discrepancy rates generated from this analysis are also understated because the discrepancy rate uses the maximum value in the denominator in order that discrepancy rates do not exceed 100%.

Another way of looking at these results is to see how well the data match between state databases and SDWIS/FED. For example, TCR MCL data have an estimated minimum discrepancy rate of 15%; this means that a maximum of 85% of the data match. Maximum correlation estimates are shown below:

TCR		Chem		SWTR		LCR			Total		Overall
MCL	M/R	MCL	M/R	TT	M/R	TT	M/R	MCL	TT	M/R	Total
85%	69%	60%	59%	80%	62%	14%	32%	82%	74%	61%	63%

Overall, roughly 2/3 of the data in state databases and SDWIS/FED match. LCR TTs had the lowest correlation estimate of 14%.

Here is how these estimates were calculated:

Again, EPA used 1997 data, only included states that reported using their own databases, and excluded Chem M/R violations from the state with huge underreporting.

EPA separated the minimum and maximum value of each pair of data (a pair of data being, for example, the number of TCR MCL violations in state databases and the corresponding number in SDWIS/FED). Each pair of data (the data point in the state database and the corresponding value in SDWIS/FED) was sorted by the maximum and minimum value. The totals were put into the following equation:

$$\text{Minimum discrepancy rate} = \frac{\text{sum of maximum \# violations} - \text{sum of minimum \# violations}}{\text{Sum of maximum \# violations}}$$

EPA divided by the maximum number of violations to keep the discrepancy rates below 100%.

5.2.4 Main reasons cited by states for these discrepancies

Overall, the category of “data entry problems” was the most common reason given for discrepancies. This includes incomplete PWS inventories; and data submission, transfer file format, and coding problems. The category of “resource limitations” was the next most common reason for discrepancies. This includes the inability of a state system to upload data to SDWIS/FED, lack of staff and/or programmers, and no automated tracking system for a particular rule.

The following number of states citing these reasons for discrepancies, by violation type:

	MCL		**M/R**		**TT**	
Reason	Under-reporting	Over-reporting	Under-reporting	Over-reporting	Under-reporting	Over-reporting
Data Entry	14	8	23	9	9	3
Resource Limitations	7		7		11	
Regulation Interpretation Issues	3		7			
ACR Guidance Interpretation Issues	5	2	6	5		1
Late Reporting	1		4	1	2	
Automated System Generation			1			
Reason not provided	20	10	31	18	11	9

For under-reporting, M/R violations had the most discrepancies. The most frequently cited reason is data entry. This was followed by TT violations, with the most frequently cited reason being resource limitations, followed by data entry. MCL violations had the fewest discrepancies. The most frequently cited reason is data entry.

Most frequently cited reasons for discrepancies, by rule:

		Data entry	Resource limitations	ACR guidance	Regulation implementation
Chems	MCL	#1	#2		
	M/R	#1	#2		
TCR	MCL	#1			
	M/R	#1	#2		
SWTR	TT	#1	#2		
	M/R	#1	#2	#2	#2
LCR	TT		#1		
	M/R		#1		

5.3 Error reports analysis—data transfer errors

5.3.1 Background

This analysis reviewed 2 quarters of error production reports (received during the period August 1 through December 31, 1998), to look at the magnitude of, and reasons for, data transfer errors between state databases and SDWIS/FED. Eight hundred forty one (841) files were reviewed. Three hundred two (302) files were analyzed in detail to determine the error rejection rate and the error correction rate.

At the state level, the information obtained will be used to provide recommendations for corrective actions, training needs identification, and quality assurance procedures. Because the method of update, error correction, and level of effort states expend on correcting errors varies from quarter to quarter, extrapolating the errors analysis information to determine a national level rejection rate for each type of submission and/or data type was determined to be inconclusive. Additional meta-data will need to be collected in the future if more detail is desired on rejection rates.

5.3.2 Common types of errors

Of the over 800 possible error conditions which are programmed into SDWIS/FED edit criteria, only 230 occurred in the 841 files analyzed. The most common reasons are listed below:

27%	Invalid values: typos, non-permitted values, etc.
14%	Cross Edits: data rejected because a comparison between two or more attributes yielded incompatible values.
8%	Non-Existent Data: attempts to modify or delete data or records which do not exist on the database.
8%	Processing Rule: comparison between two or more attributes showed invalid combinations.
8%	Missing Registration Requirements: attempts to post a new water system without all required elements present.
7%	Content: missing values and/or missing combinations of data.
7%	SDWIS/FED bugs and software limitations.
6%	Duplicate Data: data already exists in the input file or in the database.

Eighty-two percent (82%) of the error "types" relate to data entry errors (e.g., failure to follow data entry instructions, keypunch, missing or incomplete data, or invalid values). SDWIS/FED bugs and software limitations represent 7% or less of the errors. The remaining 11% included informational messages, old FRDS conversion errors, and errors that could be either a SDWIS/FED bug or a state data entry error depending on the data submitted.

5.3.3 Main reasons cited for non-reporting during the analysis period

State resource limitation was given as the primary reason for Lead & Copper sample data not being reported during the analysis period. Three states were unable to submit action files due to major system software reprogramming or data clean-up activities. Those failing to submit any inventory data during 1998 cited major system software conversion activities or state resource limitations as the reason.

5.3.4 Rejection rates of files

Rejection rates for inventory and actions data were calculated for files submitted using the Traditional method. It was not possible to calculate comparable rejection rates using the Total Replace method because SDWIS/FED cannot identify which data in the file are being submitted for the first time. The following equation was used:

$$\text{The error rates of files using the Traditional method} = \frac{\text{\# lines in error file}}{\text{\# lines in input file}}$$

In Traditional updates, 20% of inventory data and 32% of violation and enforcement actions data are being rejected.

5.3.5 States' success in submitting correction files

Most states attempted corrective actions. When data are rejected from SDWIS/FED, states (or EPA regions acting on the state's behalf) are sent error reports indicating what data were rejected and the reason(s) for the rejection. It appears that error files having a large number and/or variety of errors were not being corrected on the first attempt. Some errors did not require correction, such as duplicate records being submitted, or intentional manipulation by the state or EPA region in order to achieve a specific result. It was not possible to accurately determine the volume of such errors. Only a quarter of all states were completely successful in resubmitting rejected data on their first attempt. Three-fourths of the states had at least a quarter of the second attempt reject. Reasons for errors remaining uncorrected include: states did not understand how to correct the original error, they chose to correct only some errors, or, as mentioned above, some errors do not require correction.

5.4 State structures analysis

Analysis of the ASDWA Management and Data Flow of States survey failed to produce any clear-cut reasons for particular state drinking water programs to have better data quality, defined as consistency between state records and SDWIS/FED. To perform the analysis, state ranking was determined by dividing the total number of discrepancies for violations between State records and SDWIS/FED by the total number of violations, and obtaining a percentage. Then staff looked to see if there was a correlation between the way a state is organized and its data quality, as measured by its discrepancy rate.

Analysis showed that both the highest and lowest ranking states had similar responses to the survey questions.

Because the analysis of the ASDWA data provided no clear-cut answers, EPA asked the Cadmus Group which has conducted data verifications in the past to select several states that it believes maintain model programs and summarize the organizational structure of those states. These states were selected regardless of any violation or discrepancy numbers present in DV reports. The number or levels of the organization do not appear to have as great an impact on data quality as does the quality of communications. Another key factor was an adequate number of trained, qualified personnel. The program and management structural components which were believed to be critical to promoting high data quality, are presented below:

- Communication: Routine, meaningful and timely communication at all levels.
- Annual PWS notification of monitoring schedules and requirements
- Automated compliance determination for monitoring requirements
- Violation notification with required corrective action instructions
- Standard operating procedures and related periodic training including: data entry, forms completion, conducting sanitary surveys, and compliance determination
- Efficient and timely method of access to water system data for all staff
- Electronic access to laboratory sample data
- Existence and use of a quality assurance program which resolves and prevents errors
- Standardized data submission format (electronic or forms) from PWS and labs
- Streamlined handling process of document and analytical result handling through compliance determination and recording violations and follow-up actions

5.5 State summaries of SDWIS/FED data quality, and recommended improvements

The last component of this project is the EPA analysis of SDWIS/FED data quality on a state-by-state basis. The resulting state summary reports will provide specific prioritized recommendations to help states improve their data quality. The individual summaries will be provided to states separately during the spring of 2000. The summaries will address the following state-specific findings.

- ACR vs. SDWIS/FED analysis findings which highlight areas of zero and non-reporting, and violation type discrepancies which are greater than 10%.
- Numeric and non-numeric findings from data verifications conducted between 1996 and 1998. Data verifications conducted during 1999 are used to clarify or support findings from other analysis areas. Strengths and areas of weakness, which impact data quality, are highlighted.

- The number of violations reported by each state in each fiscal year between 1993 and 1998, from the frozen database analysis. Violations are categorized by contaminant/rule and by water system type.
- A discussion of state management structures with recommendations for improvement including a listing of key components that promote good SDWIS/FED data quality.
- Significant findings from EPA Mid-Year and/or End-of-year Program Reviews relating to data management and SDWIS/FED data quality.
- An analysis of SDWIS/FED error reports from data submitted during August 1, 1998 through December 31, 1998.

Appendix A—Stakeholders Working Group recommendations

Recommendations were identified and evaluated during the three major phases of the Data Reliability Action Plan. The first phase involved the 3 public stakeholder meetings. The second phase involved the individual analyses that were conducted, the results of which are included in this report. The third phase was the Stakeholder Work Group review of the preliminary findings of the data verifications, error report, ACR and timeliness analyses at the September 1999 meeting and additional recommendations were suggested. All recommendations were discusses and voted on at the meeting. The following table presents the results of that vote.

# votes	**Recommendation**
19	Increase training • Provide on-site assistance to resolve state-specific data entry problems. • Provide additional compliance determination training, and data entry training for new and existing rules • Establish a multi-regional cadre of trainers (funded through either a central contract and/or with the states paying for travel).
17	Improve the data verifications audits • Include specific, prioritized, implementable recommendations. • Include the # of systems with discrepancies. • Conduct DVs for each state every 2-3 years, which will help promote and track follow-up to previous DV recommendations. • Issue DV procedures so states can perform self-audits • Review data at the water system level to correlate data in state files • Add a timeliness review • Make follow-up of DVs part of regional quarterly/annual reviews • Tighten follow-up procedures—have the EPA regional office check back with states within 6 months
15	Streamline reporting and rule complexity
15	Make error reports more user-friendly. It is currently very difficult for managers to use them to identify specific problems
11	Encourage states to notify utilities annually of compliance monitoring schedules
10	EPA should focus follow-up on poorer state/regional performers • Focus on states not reporting specific rules—should trigger a focused DV audit
7	Require electronic reporting of monitoring regulations in the future
7	Require states to issue notices to utilities for each violation
6	Require labs to report sample results directly to states electronically
6	Improve front end retrieval of SDWIS/FED data
6	EPA HQ should provide contract funds for data management technical assistance
5	Provide new resources for data management

5	Enable utilities to review their data before it is sent to SDWIS/FED • Encourage state web access • Ask trade association to communicate need for states to have additional resources to enable web access
5	Establish a multi-state cadre of state peer reviewers. • States provide travel funds • Voluntary basis
4	Focus national program guidance on M/R discrepancies • Help mitigate funds drawn to other media
3	Develop automated compliance determination mechanisms in SDWIS/STATE
3	Centralize Oracle DBA support (this recommendation applies to all states, not only those using SDWIS/STATE)
3	Establish contract funds to help states enter data on an as-needed basis
2	Provide better guidance, including data flow diagrams, when new rules are issued
1	Have EPA over-file for states which choose not to report
1	Complete the edit summary report to identify generic errors
0	Standardize data transfer mechanisms